Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 21(1): 1038–1057.
DOI: 10.3934/mbe.2024043
Received: 12 October 2023
Revised: 02 December 2023
Accepted: 07 December 2023
Published: 22 December 2023

Research article

A dual-modal dynamic contour-based method for cervical vascular ultrasound image instance segmentation

Chenkai Chang[1], Fei Qi[2,*], Chang Xu[1], Yiwei Shen[1] and Qingwu Li[1]

[1] College of Information Science and Engineering, Hohai University, Changzhou 213200, Jiangsu, China

[2] Changzhou No.2 People's Hospital, No.29 Xinglong Lane, Changzhou 213004, Jiangsu, China

* **Correspondence:** Email: czey_feiqi@163.com.

Abstract: *Objectives:* We intend to develop a dual-modal dynamic contour-based instance segmentation method that is based on carotid artery and jugular vein ultrasound and its optical flow image, then we evaluate its performance in comparison with the classic single-modal deep learning networks. *Method:* We collected 2432 carotid artery and jugular vein ultrasound images and divided them into training, validation and test dataset by the ratio of 8:1:1. We then used these ultrasound images to generate optical flow images with clearly defined contours. We also proposed a dual-stream information fusion module to fuse complementary features between different levels extracted from ultrasound and optical flow images. In addition, we proposed a learnable contour initialization method that eliminated the need for manual design of the initial contour, facilitating the rapid regression of nodes on the contour to the ground truth points. *Results:* We verified our method by using a self-built dataset of carotid artery and jugular vein ultrasound images. The quantitative metrics demonstrated a bounding box detection mean average precision of 0.814 and a mask segmentation mean average precision of 0.842. Qualitative analysis of our results showed that our method achieved smoother segmentation boundaries for blood vessels. *Conclusions:* The dual-modal network we proposed effectively utilizes the complementary features of ultrasound and optical flow images. Compared to traditional single-modal instance segmentation methods, our approach more accurately segments the carotid artery and jugular vein in ultrasound images, demonstrating its potential for reliable and precise medical image analysis.

Keywords: deep learning; dual-modal; instance segmentation; ultrasound-optical flow image; cervical vascular

Table 1. Abbreviations used in this paper.

Abbreviation	Meaning	Abbreviation	Meaning
US	Ultrasound	US-F	Ultrasound and optical flow
GT	Ground truth	CNN	Convolutional neural networks
CT	Computed tomography	MRI	Magnetic resonance imaging
RGB	Visible images	RGB-Flow	Visible and optical images
RGB-T	Visible and thermal infrared image	RGB-D	Visible and depth image
MLP	Multi-layer perceptron	AHE	Adaptive histogram equalization
GAP	Global average pooling	GMP	Global max pooling
FFN	Feed-forward network	DML	Dynamic matching loss
CPU	Central processing unit	GPU	Graphics processing unit
AP	Average precisions	TP	True positives
FP	False positives	FN	False negatives
IoU	Intersection over union	AI	Artificial Intelligence
FLOPs	Floating point operations	mAP	Mean average precision
FPS	Frames per second	DSSIFM	Dual-stream spatial information fusion module
FPN	Feature pyramid network		

1. Introduction

Medical ultrasound imaging, a secure, painless, and non-invasive technique, harnesses high-frequency sound waves to provide real-time visualisation of internal bodily structures [1]. In recent years, ultrasound-guided jugular vein cannulation has been extensively implemented in clinical practice [2]. Under ultrasound guidance, physicians can directly observe vascular anatomy and needle tip location, thereby improving cannulation success rates and reducing complication occurrences compared to the traditional approach. However, inherent limitations of ultrasound imaging itself remain, such as low resolutions and inferior imaging quality [3]. Another challenge lies in differentiating intricate imaging patterns between vasculature, musculature, and neural tissues, which risks misinterpretation. Therefore, even for clinicians with extensive expertise, serious complications such as inadvertent carotid artery puncture leading to severe hemorrhage continue to occur despite ultrasound guidance during jugular vein cannulation procedures [4]. Accurate identification of jugular venous and carotid arterial anatomy is thus imperative to improve outcomes of this intervention. To achieve such goals, many researchers have proposed adopting deep learning-based techniques [5].

Propelled by the meteoric advancement of computing technology over the recent years, deep learning has become ubiquitous across most natural scenes, including natural language processing [6], computer vision [7] and financial domains [8, 9]. Essentially, deep learning is a method of using a large amount of known data to fit a nonlinear model that can be used to predict unknown situations [10]. The field of medical image processing [11] is also endeavoring to adopt deep learning-based approaches, in an effort to enhance the efficiency and accuracy of medical image analysis. U-Net [12] showed good segmentation results due to the use of encoder-decoder structure and skip connections. Subsequently, many scholars improved U-Net. Zhou et al. [13] proposed U-NET++, which incorporates a series of nested and dense convolutional blocks in the skip connections to reduce semantic gaps between feature maps. The DDAUnet proposed by Yousefi et al. [14] utilized spatial and channel attention gates in each dense block to selectively focus on features and regions of interest, enabling automatic delineation of tumor tissue in esophageal CT images. Swin-Unet [15]

constructed a U-shaped encoder-decoder architecture based on Transformer modules, utilizing self-attention mechanisms to learn semantic representations from local to global contexts, enabling medical image segmentation.

However, these existing methods focus primarily on fusing or connecting feature maps at different levels extracted by the backbone networks. They neglect intrinsic issues of the original ultrasound images themselves, such as poor imaging quality, indistinct boundaries between organ tissues, and failure to exploit inter-frame information across consecutive ultrasound image sequences [16]. As shown in Figure 1, the red box outlines the carotid artery and jugular vein area, but its texture structure is highly similar to the nearby muscle tissue (yellow dotted box). In CNN architectures, segmenting areas with analogous textures are a formidable challenge [17].

To address the inherent deficiencies of data from a single modality, some researchers have begun exploring the fusion of dual-modal information to enhance the performance of image processing. In the context of RGB-T (visible and thermal infrared image) [18] image segmentation, the complementary information from visible light and thermal infrared images has been utilized to improve segmentation accuracy. Thermal infrared images provide temperature information of a scene but lack visual details like color and texture. Conversely, RGB (visible) images contain color, texture, and other visual information, which are essential for identifying the categories and appearances of objects. Similarly, the use of motion information from optical flow images as auxiliary supervision has been proven to enhance segmentation performance in natural scenes [19].

Optical flow represents the motion or displacement of each pixel between consecutive frames in an image sequence, offering additional details about motion and contours [20, 21]. It can capture vascular pulsation movements and geometric structural changes that are invisible in static ultrasound images, thus significantly enhancing the accuracy of ultrasound image instance segmentation. Inspired by the achievements in dual-modal image processing, we propose to integrate motion and contour information from optical flow images with ultrasound images for the instance segmentation of carotid artery and jugular vein. This approach not only fully utilizes the tissue information inherent in ultrasound images but also leverages the additional motion and boundary details provided by optical flow images, which is beneficial to improving the clarity and precision of segmentation.

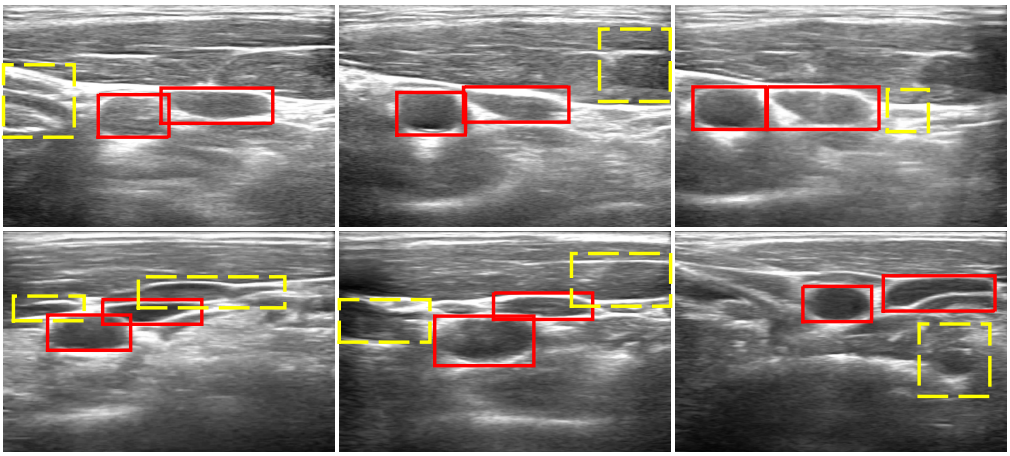

Figure 1. Subset of ultrasound image dataset.

The main contributions of this work are summarized as follows:

1) We extract optical flow information from consecutive frames of ultrasound images, and propose a dual-modal instance segmentation method for carotid artery and jugular vein. By exploiting inter-frame vascular motion information, our approach aims to improve the segmentation performance for carotid artery and jugular vein;

2) We design a dual-stream spatial information fusion module to fuse four layers of multi-level features extracted from the ultrasound image backbone and the optical flow image backbone based on the spatial attention mechanism, making full use of the complementary information between the two modal images;

3) We propose a Transformer-base global contour point deformation module that utilizes self-attention and cross-attention mechanisms. This module learns contextual representations of each point relative to all other points on the contour, enabling swift coordinated movement of all contour points to fit the ground truth points and obtaining smoother edges of the vascular segmentation;

4) We evaluate our method with other methods on a self-built carotid artery and jugular vein instance segmentation dataset. The segmentation metrics of our method are ahead of the vast majority of other methods, and the qualitative experiments show that our method can achieve smoother blood vessel edges.

2. Related work

2.1. Natural scene image segmentation

Common natural scene image modalities include RGB (visible) images, depth images, thermal infrared images and hyperspectral images. Each modality provides distinct information about the visual scenes—RGB (visible) images capture color and texture cues, depth images provide 3D structural information, thermal images reveal heat signatures, and hyperspectral data encodes rich spectral characteristics.

2.1.1. Single-modal image segmentation

The essence of single-modal image processing is to use only images obtained by one sensor to implement image processing technology. Most scholars are committed to researching the segmentation of RGB (visible) images. The fully convolutional network (FCN) [22] network proposed by Long et al. can accept inputs of any size and produce outputs of corresponding sizes through effective reasoning and learning, which has become the pioneering work of image semantic segmentation. Rafique et al. [23] performed indoor/outdoor scene detection and classification and statistical multi-object segmentation through depth images. Civilibal et al. [24] used deep learning methods to segment breast lesions from thermal infrared images of human breasts. Yu et al. [25] proposed a cross-level spectral-spatial joint encoding framework for imbalanced hyperspectral image segmentation.

2.1.2. Dual-modal image segmentation

Nowadays, many inexpensive sensors are widely available. Through extensive experiments, researchers have discovered that if a model utilizes not only features from a single data modality, but

integrates features from different data modalities, it can make the model more robust and accurate when facing various complex situations. The common dual-modal data include RGB-Flow (visible and optical flow images), RGB-T (visible and thermal infrared images), RGB-D (visible and depth images). Al-Battal et al. [26] proposed to simultaneously utilize ultrasound and optical flow video information for real-time tracking of anatomical structures. Xu et al. [27] proposed a Dual-space graph-based interaction network to capture long-range dependencies between RGB (visible images) and thermal imaging modalities for cross-modal fusion, enabling all-weather semantic segmentation of power equipment in high-voltage transmission line and substation scenarios. Sun et al. [28] proposed a cascaded transformer decoder with multi-level feature fusion and feedback to enhance and seamlessly aggregate dual-modal multi-scale contexts for RGB-D (visible and depth images) salient object detection.

2.2. Medical image segmentation

Medical images visualize the anatomical structures and functional tissues of the human body, mainly through imaging techniques such as computed tomography (CT), magnetic resonance imaging (MRI), X-ray and Ultrasound. Li et al. [29] combined the EresUNet++ architecture, residual blocks and efficient channel attention modules to design an algorithm for segmenting liver tumors in liver CT images. Raza et al. [30] proposed an end-to-end framework for automatic 3D brain tumor segmentation based on a hybrid of deep residual networks and U-Net models. Hou et al. [31] designed a network for tooth X-ray segmentation based on contextual semantic information and contrast enhancement to solve the problem of blurred boundaries between teeth in X-ray images. Yang et al. [32] connected CNN and Swin transformer as the backbone feature extraction networks, and proposed a method for automatic segmentation of breast tumors in breast ultrasound images based on channel attention to enhance important feature regions.

3. Materials and methods

3.1. Ultrasound dataset collection

The dataset consists of high-quality ultrasound images collected from 10 volunteers using a state-of-the-art 'Lumify' portable handheld ultrasound device [33] manufactured by Philips with an ultrasound probe (model L12-4). The collection process follows strict protocols to ensure the accuracy and consistency of data collection. To perform carotid artery and jugular vein ultrasound examinations, participants were in supine position with the neck extended backwards. This allowed optimal exposure of the carotid artery and jugular vein to the skin surface for ultrasound scanning. Prior to formal examination, participants rested in this position for at least 5 minutes to allow haemodynamic conditions to stabilize. In addition, room temperature was kept constant and all participants used the same brand of ultrasound gel. This minimized potential effects of different gels on image quality. By adopting consistent examination preparation and operation procedures, variability could be reduced to the largest extent possible.

To further ensure the reliability and robustness of the dataset, we selected the ultrasound images of five adult males and five adult females aged 20–25 years. An ultrasound specialist oversaw the image acquisition process, minimising potential variations and guaranteeing the accuracy of the annotations.

Guided by a team of skilled medical professionals, each ultrasound image is thoroughly annotated using Labelme [34] software to accurately outline the areas of interest associated with the carotid artery and jugular vein in ultrasound images. The total size of the dataset comprised 2432 ultrasound images and 2432 annotated images with case information annotations. We divided 2432 ultrasound images into a training dataset, a validation dataset and a test dataset, in which the training dataset contained 1950 images, while the validation and test dataset contained 241 images each.

3.2. Optical flow dataset generation

Effective image preprocessing is crucial for enhancing the quality of ultrasound images and reducing the errors caused by inherent noise and artefacts. Ultrasound imaging is susceptible to speckle noise and phantom artefacts because of the imaging principles involved. Prior to generating optical flow data from ultrasound image sequences, we applied speckle-reducing anisotropic diffusion denoising techniques to suppress the noise present in the ultrasound images. This method effectively reduces the speckle noise while preserving important image features.

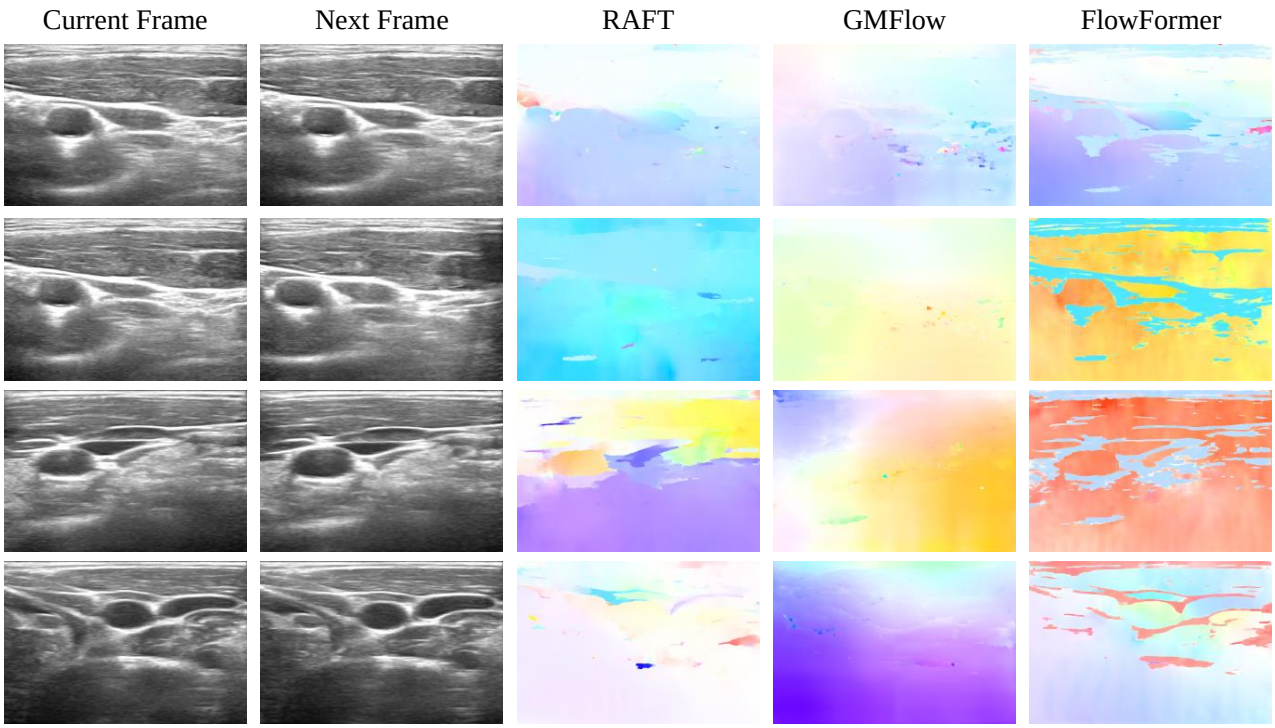

Figure 2. Comparison of visual results of three optical flow networks.

In addition, to mitigate the common intensity inhomogeneity issue in ultrasound images, we utilized adaptive histogram equalization (AHE) [35]. AHE divides an image into contextual regions and applies contrast enhancement to each region. By adjusting the local contrast within regions of the image, finer anatomical structures and boundaries can be emphasised, while the interference from intensity inhomogeneities is reduced.

We used the three latest and most advanced optical flow algorithms—GMFlow [36], RAFT [37],

and FlowFormer [38]—to generate dense optical flow datasets from ultrasound image sequences of our self-built dataset. Optical flow depicts the contours and spatial information of the carotid artery and jugular vein by capturing pixel-by-pixel motion changes between consecutive frames in the ultrasound image dataset. The optical flow visualisation results generated by the three algorithms for the ultrasound image pairs are shown in Figure 2. After a qualitative comparison of the optical flow outputs, we chose the FlowFormer [38] network as the best algorithm for generating ultrasound optical flow datasets, as it can accurately capture complex motions while maintaining a smooth flow field.

3.3. Model architecture

In this section, we present a thorough explanation of our innovative network architecture, designed specifically for the segmentation of carotid artery and jugular vein within ultrasound images.

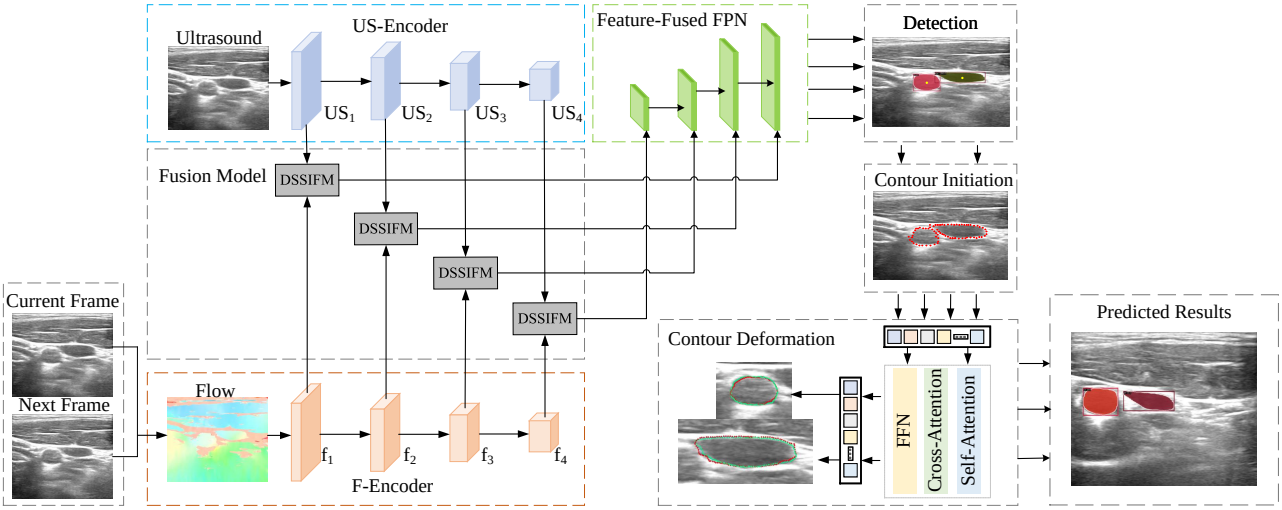

Figure 3. The overall architecture of the proposed network.

First, a concise overview of the architectural framework is provided in Figure 3. Subsequently, we comprehensively elucidate the distinct stages that constitute our ultrasound and optical flow feature fusion pyramid module. This module is a pivotal component that efficiently amalgamates the characteristic information from these two disparate modes.

Furthermore, we briefly introduce the functionality of the object detection module. This module serves as the foundational element preceding our instance segmentation task, which is tasked with obtaining the initial centre coordinate position of the object during the contour initialisation phase. Equally vital, the contour initialisation and refinement modules assume crucial roles in meticulously defining refined object boundaries. These modules leverage advanced techniques to enhance the segmentation precision.

Through a meticulous presentation of the proposed network architecture and its encompassing suite of modules, our objective is to foster a comprehensive understanding of our approach to instance segmentation in carotid artery and jugular vein ultrasound images. This methodology synergistically employs advanced techniques, including feature fusion, target detection, and contour initialisation

refinement, to obtain precise and intricate segmentation results.

3.3.1. Overall architecture

By combining contour-based optimisation with dual supervised learning, our approach effectively leverages both local and global contextual cues. This dual-branch framework facilitates better feature representation and enables the model to capture intricate details, leading to significantly improved instance segmentation results.

Our approach extracts multilevel ultrasound and optical flow features using distinct encoders. In the case of the ultrasound encoder, the input image US (ultrasound) is processed using the ResNet-50 [39] backbone, yielding a set of multi-scale features denoted as $(US_i, i = 1, ..., 4)$. Simultaneously, the optical flow stream employs a CNN backbone that mirrors the structure of the ultrasound encoder to derive its own multiscale features, denoted as $(f_i, i = 1, ..., 4)$, with independent weights. The method proposed in this paper can supplement the cross-modal spatial information of feature maps extracted from the ultrasound and optical flow image feature backbones. This is achieved using the proposed dual-stream spatial information fusion module. The fusion features are then integrated into a feature pyramid, resulting in a rich variety of multi-scale features.

The object detection process commences with the application of the multi-head FCOS [40] method, yielding detection boxes. Subsequently, the segmentation masks corresponding to each target box are extracted. Notably, the target contour originates from the features of the detection box centres. Subsequent refinement utilises an MLP to optimise the initial contour points. Building on this foundation, we propose a contour optimisation technique centred on a transformer network. This method involves obtaining offsets through the self-attention mechanism of the contour and its cross-attention with other features. Iteratively, finely detailed contour edges emerge, culminating in the successful instance segmentation of the ultrasound images.

3.4. Dual-stream spatial information fusion module

Ultrasound images provide rich information on visual appearance and texture details, whereas optical flow maps contain more prior knowledge of motion and provide rich vascular boundary information. The complementary information between the ultrasound flow and optical flow maps plays a crucial role in the instance segmentation of US-F. The efficient fusion of complementary information between these two modal features will be committed to achieve accurate segmentation of carotid artery and jugular vein ultrasound images. Low-level feature maps typically contain more spatial information because of their high resolution, whereas high-level features typically contain more semantic information. Based on this, we propose a module based on dual-stream spatial information attention fusion, which interactively fuses the high- and low-level information of the two modalities, further improving the performance of US-F instance segmentation.

The dual-stream spatial information fusion module (DSSIFM) enhances the complementary information of low-level features of ultrasound and optical flow modes based on the spatial attention mechanism. The complete DSSIFM architecture is shown in Figure 4.

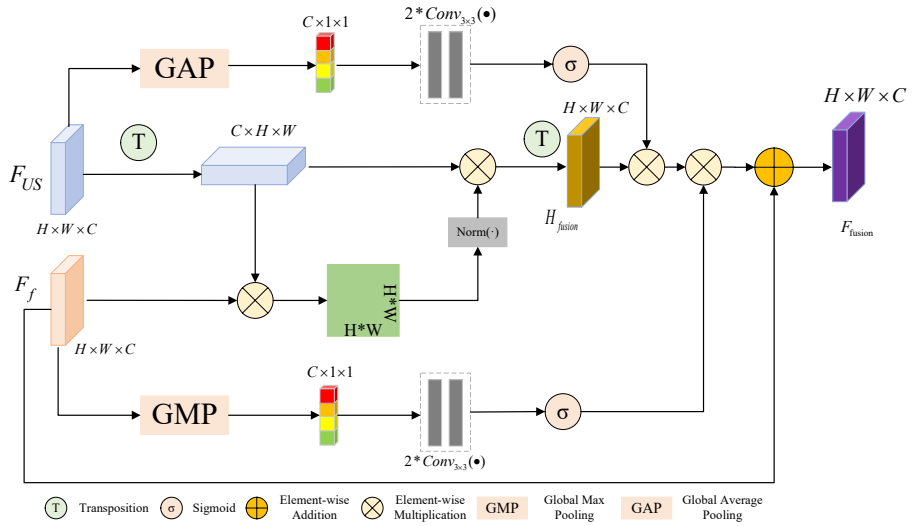

Figure 4. Dual stream spatial information fusion module.

Specifically, we obtain the ultrasound image features F_i^{US} and optical flow features F_i^f from their respective backbone networks, where $i \in \{1, 2, 3, 4\}$ indicates the stage indices of the encoder. First, advanced fusion feature representations are extracted to learn the spatial relationship between F_i^{US} and F_i^f.

$$H_i^{fusion} = Norm\left(\left(F_i^{US}\right)^T \otimes F_i^f\right) \otimes \left(F_i^{US}\right)^T. \tag{3.1}$$

where $Norm(\bullet)$ denotes $l-2$ normalisation, T stands for the transposition operation, and $\otimes$ represents the element-wise multiplication operation.

Optical flow images provide additional contour dynamic instance information to augment the proposed method. To integrate complementary insights from the ultrasound and optical flow modalities more effectively, we have designed specialised global pooling operations for each stream. For the ultrasound backbone features, we apply global average pooling (GAP) to aggregate spatial information and obtain holistic representations encoding the overall semantic content. For the optical flow stream, we utilised global max pooling (GMP) to highlight the most discriminative motion patterns and preserve the localised cues. Using this tailored pooling strategy, holistic image-level semantics and fine-grained motion signatures are synergistically incorporated. These modalities are adaptively translated into a unified representation enriched by the complementary strengths of the two information sources. GAP provides a stabilising effect to capture the overall gist, whereas GMP selectively accentuates informative motion dynamics. In summary, we use two different pooling operations and two 3×3 convolution selected rows for the ultrasound stream and optical flow stream, respectively, to integrate the complementary information of the two modes, which can be expressed as:

$$F_i^{US(\Delta)} = \sigma\left(Conv_{3\times3}\left(Conv_{3\times3}\left(GAP\left(F_i^{US}\right)\right)\right)\right), \tag{3.2}$$

$$F_i^{f(\Delta)} = \sigma\left(Conv_{3\times3}\left(Conv_{3\times3}\left(GMP\left(F_i^{f}\right)\right)\right)\right). \tag{3.3}$$

To retain the holistic semantics from the ultrasound stream, we have incorporated a residual connection

before fusing the modalities. The final fused feature representation is formulated as follows:

$$F_i^{fus} = \left(\left(H_i^{Fusion} \otimes F_i^{f(\Delta)}\right) \otimes F_i^{US(\Delta)}\right) \oplus F_i^{US}. \tag{3.4}$$

3.5. Transformer-base contour deformation

After passing through our DSSIFM, the four fused features are fed into a feature pyramid network (FPN) [41] for extraction. The FPN [41] enables the network to assemble semantically stronger features by leveraging the inherent multi-scale pyramidal hierarchy of deep convolutional networks. Specifically, we utilize a top-down architecture with lateral connections to build an in-network feature pyramid from a single-scale input. The FPN [41] considers the four fused features as inputs and generates a feature pyramid with semantics from low to high levels. Using this robust feature representation, our model can effectively learn the global structure and local details for accurate carotid artery and jugular vein segmentation in ultrasound images.

3.5.1. Object detection module

In the context of the object detection task, we have opted to utilise the FCOS [40] detector, which is seamlessly integrated into the network architecture. Specifically, multiple object detection heads within the FCOS [40] framework have been employed to facilitate multi-scale predictions. In contrast to anchor-based approaches such as YOLO [42] and RCNN [43], FCOS [40] object detection heads directly infer vectors for classification labels with dimensions corresponding to the number of classes, along with vectors for bounding box coordinates with dimensions four times that of the class count.

3.5.2. Contour initialization and refinement

Current contour evolution methods are limited because they lack global context modelling. Each contour point leverages only local neighbouring information and is unable to exploit the long-range dependencies along the contour. To address this issue, we propose a Transformer-base approach that enables interactions between all the contour points.

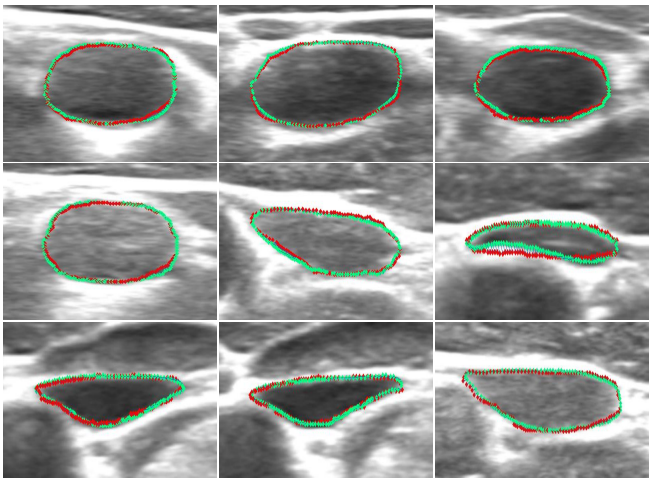

Figure 5. Demonstration process of contour initialization and refinement (The red dots represent the initial point of the contour, and the green dots represent the refined contour).

Specifically, we modify the object detection transformers DETR [44] and Mask2Former [45] for contour optimisation. These provide encoder-decoder architectures that are suitable for our task. In the encoder, we extract the regional features of the entire image to enrich the global context. The decoder then predicts the contour point offsets in parallel. A key consideration is the initialisation of contour point queries for the decoder. Unlike DETR [44], which uses learned queries, we initialise them directly from the centre-point features. This provides a strong starting point for each image. Next, self-attention in the decoder models the relationships between all the contour points. This allows each point to gather the context from the entire contour. Furthermore, the cross-attention between the points and encoder region features incorporates the global image context. An object's spatial features specifically serve as keys to cross-attention. For offset prediction, a feed-forward network (FFN) regresses to the x and y offsets for each decoded point. The FFN uses standard transformer layers to process attentive point features. Finally, by decoding the point features enhanced by the global context, the FFN can predict precise contour offsets in an end-to-end manner. Figure 5 shows the results of contour initialization and refinement modules proposed by us for fitting the boundaries of carotid artery and jugular vein. With the modules we proposed, smoother and clearer carotid artery and jugular vein contour edges can be obtained.

3.6. Loss function

Our proposed method sets two loss functions: Smooth L1 loss and dynamic matching loss. Smooth L1 loss helps eliminate large deviations between the predicted profile and the real profile, bringing the profile closer to the target boundary. Dynamic matching loss provides more flexible matching between the predicted and target points along the profile. These losses are defined as follows:

$$L_{init} = \frac{1}{N}\sum_{i=1}^{N} smoothl1\left(\tilde{x}_i^{init} - x_i^{gt}\right), \tag{3.5}$$

$$L_{coarse} = \frac{1}{N}\sum_{i=1}^{N} smoothl1\left(\tilde{x}_i^{coarse} - x_i^{gt}\right), \tag{3.6}$$

$$L_{iter1} = \frac{1}{N}\sum_{i=1}^{N} smoothl1\left(\tilde{x}_i^{iter1}, x_i^{gt}\right), \tag{3.7}$$

where N represents the number of contour vertices, $\tilde{x}_i^{init}$ denotes the predicted initial contour vertex, $\tilde{x}_i^{coarse}$ signifies the predicted coarse contour vertex, x_i^{gt} represents the label contour vertex, and $\tilde{x}_i^{iter1}$ characterizes the post-deformation contour vertex through the initial refinement module. For the dynamic matching loss function, we use the same DML function as E2EC [46]. The loss function for contour refinement deformation is as follows:

$$L_{iter2} = L_{DML}\left(\tilde{x}_i^{iter2}, x_i^{gt}\right), \tag{3.8}$$

$$L_{iter} = L_{iter1} + L_{iter2}, \tag{3.9}$$

$$L_{overall} = L_{\text{det}} + L_{init} + L_{coarse} + L_{iter}. \tag{3.10}$$

L_{det} is the loss of the center point detection.

3.7. Model training

The experiments were conducted on a workstation with an Intel Core i7-13700 CPU and NVIDIA RTX 4090 GPU. We implement our proposed network using the Pytorch and MMDetection [47] frameworks, and the ultrasound and optical flow images are both resized to 256 × 256 pixels for input. The training regimen spanned 50 epochs and used a single NVIDIA RTX 4090 GPU. For the optimisation, we employed the Adam optimiser with an initial learning rate of 1e-4. The training batch size was set to 8. To govern learning rate scheduling, we adopted a polylearning rate policy with a power value of 0.9.

3.8. Quantitative evaluation metrics

The average precisions (APs) for the detection boxes and segmentation masks are calculated to evaluate the proposed method.

The quality of bounding box prediction and mask prediction is evaluated in terms of standard AP metrics and Intersection over Union (IoU). To distinguish between the standard AP, IoU, and other metrics, they are denoted as AP_{mk} and IoU_{mk}, respectively. As shown in Eq (3.11), IoU_{mk} represents the ratio of the intersection and union between the mask predicted by instance segmentation and the GT mask.

$$IoU_{mk} = \frac{m_{pre} \cap m_{gt}}{m_{pre} \cup m_{gt}}, \tag{3.11}$$

where m_{pre} is the mask predicted by instance segmentation and m_{gt} is the ground truth mask. The precision and recall are calculated using instance segmentation results based on a specific IoU.

$$Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN}, \tag{3.12}$$

where TP represents the correct classification or the result where the IoU of m_{pre} and m_{gt} is greater than the preset threshold of IoU, FP represents the instance that does not exist or the result where the IoU of m_{pre} and m_{gt} is less than the preset threshold of IoU, and FN represents the missing detection. The AP of the prediction result is defined as:

$$AP = \int_0^1 P(r)dr. \tag{3.13}$$

where P is the precision and r is the recall. AP_{mk} represents the weighted average result of the IoU thresholds calculated in the range [0.5, 0.95] in steps of 0.05. In addition, the AP of a single fixed threshold is used for evaluation, such as AP_{50} (threshold of IoU is 0.5) and AP_{75} (the threshold of IoU is 0.75). For objects of different sizes, the proposed study sets different AP representations, such as AP_s (small objects with an area less than 64^2 pixels), AP_M (medium objects with an area greater than 64^2 pixels and less than 96^2 pixels), AP_L (large objects with an area greater than 96^2 pixels).

4. Results

In this section, we compare the performance of our proposed method with other networks for carotid artery and jugular vein segmentation using our self-built ultrasound dataset.

4.1. Quantitative experimental analysis

Table 2. Comparison of carotid artery and jugular vein segmentation performance for different networks (best results are shown in bold).

Model	Bounding box			Mask segmentation			FLOPs↓	Params↓	FPS↑
	mAP↑	AP50↑	AP75↑	mAP↑	AP50↑	AP75↑			
Mask R-CNN [48]	0.77	**0.953**	0.93	0.836	0.943	**0.935**	447.71	44.17	12.9
YOLACT [49]	0.795	0.877	0.873	0.803	0.877	0.874	403.62	35.29	23.7
CondInst [50]	0.786	0.942	0.929	0.707	0.931	0.801	332.8	34.16	25.3
SOLO [51]	-	-	-	0.692	0.736	0.671	358.4	36.30	22.5
PointRend [52]	0.755	0.872	0.893	0.787	0.857	0.891	299.68	56.21	16.0
SOLOV2 [53]	-	-	-	0.714	0.876	0.780	335.7	46.23	21.4
QueryInst [54]	0.722	0.932	0.917	0.823	0.942	0.933	621.42	172.46	7.6
Mask2Former [45]	0.764	0.625	0.751	0.727	0.913	0.777	248.83	44.02	23.1
BoxInst [55]	0.711	0.889	0.783	0.756	0.882	0.746	430.82	53.95	17.3
SparseInst [56]	-	-	-	0.742	0.874	0.753	268.6	37.62	21.6
RTMDet [57]	0.793	0.913	0.825	0.748	0.854	0.811	343.9	35.9	23.2
Ours	**0.814**	0.942	**0.932**	**0.842**	**0.944**	0.932	**212.4**	**32.27**	**31.5**

Specifically, we evaluated Mask R-CNN [48], YOLACT [49], CondInst [50], SOLO [51], PointRend [52], SOLOV2 [53], QueryInst [54], Mask2Former [45], BoxInst [55], SparseInst [56], RTMDet [57] and our proposed network. We utilized the self-built ultrasound image dataset to quantitatively evaluate the performance of our proposed network against other networks for carotid artery and jugular vein ultrasound image segmentation. Our evaluation encompasses several key performance indicators, including mAP (mean average precision), AP50 (Average Precision at 50% Intersection over Union), and AP75 (Average Precision at 75% Intersection over Union). Furthermore, computational complexity and efficiency indicators, such as FLOPs (Floating point operations), number of parameters, and FPS (Frames per second), were also considered. The quantitative results are shown in Table 2. Our proposed network achieved a bounding box detection mAP of 0.814, surpassing several popular models such as Mask R-CNN (0.77), YOLACT (0.795), CondInst (0.786), and RTMDet (0.793). This underscores our network's accuracy in identifying the structures of the carotid artery and jugular vein. Under the indicators of the AP50 and AP75, our proposed network continued its exemplary performance, reaching 0.942 and 0.932, respectively, further validating the efficacy of our approach. In terms of mask segmentation, our network also

demonstrated exceptional performance, with a mAP of 0.842, significantly higher than many other networks (e.g., PointRend at 0.787, QueryInst at 0.823). Although the AP50 for bounding box and the AP75 for mask segmentation of Mask R-CNN slightly exceed those of our proposed network, the Mask R-CNN network requires more parameters and higher computational power than our network. Notably, our network demonstrates superior performance in the FPS (frames per second) indicator, reflecting its greater efficiency. In summary, our network not only exhibited outstanding accuracy and robustness in the segmentation task of carotid artery and jugular vein ultrasound images but also demonstrated superior computational efficiency.

4.2. Qualitative experimental analysis

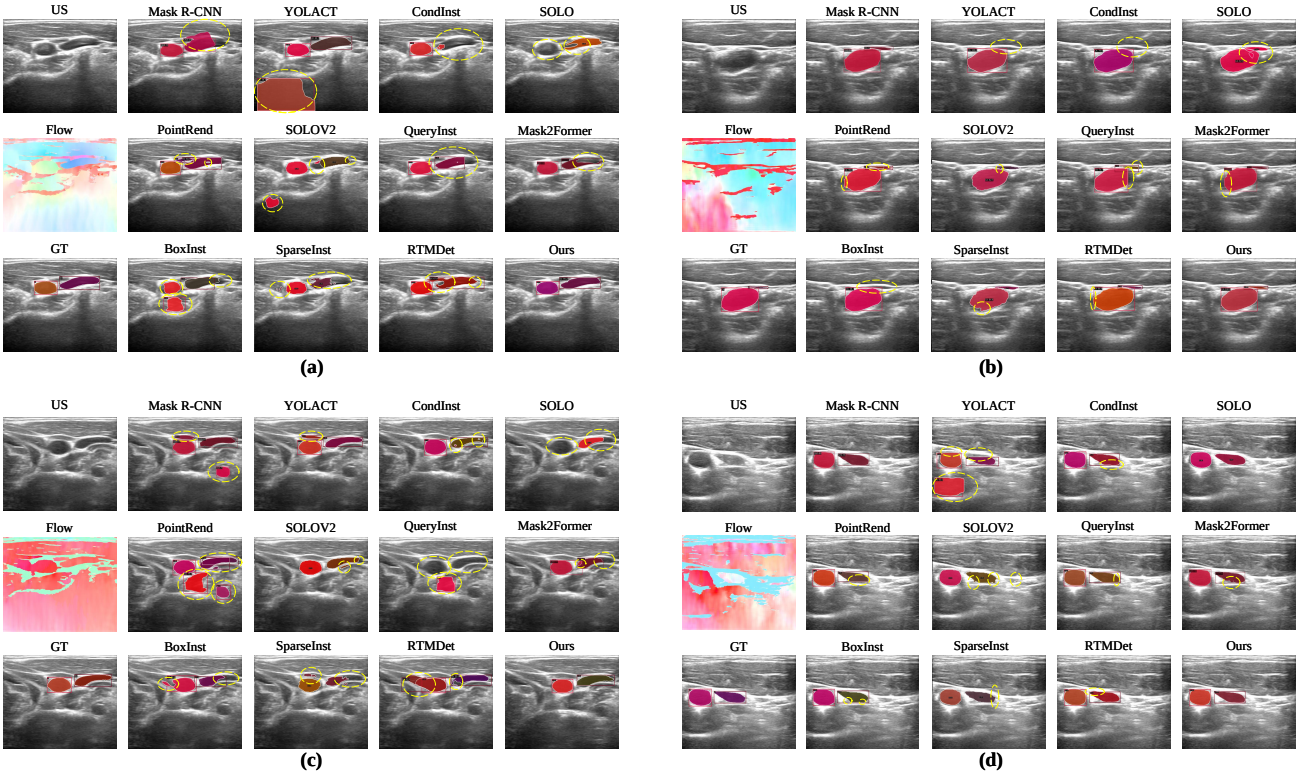

Figure 6. Comparison of experimental results for different networks.

As illustrated in the qualitative analysis of the segmentation results in Figure 6, our method can produce smoother segmentation results with clearer boundaries between the carotid artery and jugular vein, demonstrating the superiority of the proposed approach. By contrast, other networks generate segmentation with more discontinuities and indistinct boundaries. In summary, the experiments verify the effectiveness of our proposed method in accurately delineating the carotid artery and jugular vein vessels in ultrasound images. Figure 6 highlights the defective areas in the current segmentation approach, as indicated by yellow circles. Compared to other networks, our algorithm demonstrates noticeable improvements in segmenting the carotid artery and jugular vein. Specifically: 1) Other algorithms exhibit overlapping vessel segmentations, whereas our algorithm produces a clearer delineation; 2) Some vascular structures are missed by the other algorithms, and our algorithm

successfully segments them; and 3) Other algorithms may falsely identify nonvascular tissues as vessels, an issue mitigated by our approach. These advantages stem largely from the integration of the optical flow image features into the network architecture. This enhances the capability of the network to learn vascular motion contours, thereby enabling a more precise segmentation of the vascular anatomy.

5. Discussion

Medical image segmentation, whether it is X-ray images, ultrasound images, or CT images, has always been a task that the medical and computer vision communities are committed to solving. For manual detection in the traditional sense or simple image processing, such as threshold segmentation, it is difficult to completely guarantee the accuracy and stability of segmentation. Introducing deep learning methods into medical image segmentation is a great initiative. The birth of the U-Net [12] network has promoted the development of medical image segmentation. However, most medical image segmentation networks are based on single-modal image input, and the inherent shortcomings of images are still retained. In this research endeavour, we have successfully introduced a pioneering instance segmentation approach tailored to the nuanced task of carotid artery and jugular vein segmentation. By ingeniously incorporating optical flow information alongside traditional ultrasound information inputs, our proposed method has demonstrated exceptional proficiency, elevating the precision and robustness of segmenting these intricate vascular structures.

The integration of optical flow data has significantly bolstered the capabilities of our model by capturing motion status information that harmonises with the spatial cues offered by the ultrasound images. This strategic combination not only enhances the differentiation between carotid artery and jugular vein but also enriches the overall contextual understanding, resulting in superior segmentation outcomes. The utility of optical flow data as a supervisory signal reflects its efficacy in guiding the network for more accurate and reliable segmentation.

Our innovation extends further with the Dual-stream spatial information fusion module, which strategically amalgamates features from both the ultrasound and optical flow domains. The hierarchical fusion mechanism of this module facilitates the comprehensive incorporation of multi-modal information, effectively addressing the challenge of harnessing the distinctive characteristics from different data sources. The resultant enhancement in feature representation significantly contributes to the overall performance boost observed in our method. The experimental results clearly illustrate the advantages of our proposed approach, outperforming existing methods and setting new benchmarks for carotid artery and jugular vein segmentation. Beyond its specific applications, our approach lays the foundation for broader implications in medical image analysis. The seamless integration of optical-flow-derived motion status information into an instance segmentation framework holds promise for similar segmentation tasks, thereby extending its potential impact across diverse medical imaging domains.

6. Conclusions

Our quantitative and qualitative experimental results demonstrate the effectiveness of the proposed dual-modal instance segmentation network. By simultaneously utilizing the differential and

complementary information between ultrasound and optical flow images, we successfully improve the segmentation performance for the carotid artery and jugular vein in ultrasound images, achieving smoother vessel segmentation results and clearer boundaries between carotid artery and jugular vein. Although our method achieves good performance, generating optical flow images requires additional computational resources. In the future, we will devote efforts to studying real-time optical flow image generation, conducting experiments on more medical image datasets, and promoting the development of multi-modal technology in medical image segmentation.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by Youth Talent Support Project of Changzhou Health Commission (CZQM2023014), and the Key Research and Development Plan of Jiangsu Province (BE2020092)

Conflict of interest

The authors declare there is no conflict of interest.

References

1. S. Wang, M. E. Celebi, Y. D. Zhang, X. Yu, S. Lu, X. Yao, et al., Advances in data preprocessing for biomedical data fusion: An overview of the methods, challenges, and prospects, *Inf. Fusion*, **76** (2021), 376–421. https://doi.org/10.1016/j.inffus.2021.07.001
2. C. Fournil, N. Boulet, S. Bastide, B. Louart, A. Ambert, C. Boutin, et al., High success rates of ultrasound-guided distal internal jugular vein and axillary vein approaches for central venous catheterization: A randomized controlled open-label pilot trial, *J. Clin. Ultrasound*, **51** (2023), 158–166. https://doi.org/10.1002/jcu.23383
3. W. Choi, B. Park, S. Choi, D. Oh, J. Kim, C. Kim, Recent advances in contrast-enhanced photoacoustic imaging: Overcoming the physical and practical challenges, *Chem. Rev.*, **123** (2023), 7379–7419. https://doi.org/10.1021/acs.chemrev.2c00627
4. L. Wang, J. Bai, J. Jin, K. Zhi, S. Nie, L. Qu, Treatment of inadvertent cervical arterial catheterization: Single-center experience, *Vascular*, **31** (2023), 791–798. https://doi.org/10.1177/17085381221083161
5. L. A. Groves, B. VanBerlo, N. Veinberg, A. Alboog, T. M. Peters, E. C. Chen, Automatic segmentation of the carotid artery and internal jugular vein from 2D ultrasound images for 3D vascular reconstruction, *Int. J. Comput. Assisted Radiol. Surg.*, **15** (2020), 1835–1846. https://doi.org/10.1007/s11548-020-02248-2
6. D. Khurana, A. Koli, K. Khatter, S. Singh, Natural language processing: State of the art, current trends and challenges, *Multimedia Tools Appl.*, **82** (2023), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

7. C. Li, X. Li, M. Chen, X. Sun, Deep learning and image recognition, in *2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT)*, (2023), 557–562. https://doi.org/10.1109/ICEICT57916.2023.10245041
8. T. Jin, H. Xia, Lookback option pricing models based on the uncertain fractional-order differential equation with Caputo type, *J. Ambient Intell. Hum. Comput.*, **14** (2023), 6435–6448. https://doi.org/10.1007/s12652-021-03516-y
9. T. Jin, X. Yang, Monotonicity theorem for the uncertain fractional differential equation and application to uncertain financial market, *Math. Comput. Simul.*, **190** (2021), 203–221. https://doi.org/10.1016/j.matcom.2021.05.018
10. N. Shlezinger, J. Whang, Y. C. Eldar, A. G. Dimakis, Model-based deep learning, *Proc. IEEE*, **111** (2023), 465–499. https://doi.org/10.1109/JPROC.2023.3247480
11. S. Suganyadevi, V. Seethalakshmi, K. Balasamy, A review on deep learning in medical image analysis, *Int. J. Multimedia Inf. Retr.*, **11** (2022), 19–38. https://doi.org/10.1007/s13735-021-00218-1
12. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
13. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, UNet++: Redesigning skip connections to exploit multiscale features in image segmentation, *IEEE Trans. Med. Imaging*, **39** (2020), 1856–1867. https://doi.org/10.1109/TMI.2019.2959609
14. S. Yousefi, H. Sokooti, M. S. Elmahdy, I. M. Lips, M. T. M. Shalmani, R. T. Zinkstok, et al., Esophageal tumor segmentation in CT images using a dilated dense attention unet (DDAUnet), *IEEE Access*, **9** (2021), 99235–99248. https://doi.org/10.1109/ACCESS.2021.3096270
15. H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, et al., Swin-unet: Unet-like pure transformer for medical image segmentation, in *Computer Vision–ECCV 2022 Workshops*, Springer, (2023), 205–218. https://doi.org/10.1007/978-3-031-25066-8_9
16. T. S. Mathai, V. Gorantla, J. Galeotti, Segmentation of vessels in ultra high frequency ultrasound sequences using contextual memory, in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019*, Springer, (2019), 173–181. https://doi.org/10.1007/978-3-030-32245-8_20
17. R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, W. Brendel, ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, *arXiv preprint*, (2022), arXiv:1811.12231. https://doi.org/10.48550/arXiv.1811.12231
18. K. Song, Y. Zhao, L. Huang, Y. Yan, Q. Meng, RGB-T image analysis technology and application: A survey, *Eng. Appl. Artif. Intell.*, **120** (2023), 105919. https://doi.org/10.1016/j.engappai.2023.105919
19. X. Zhang, A. Boularias, Optical flow boosts unsupervised localization and segmentation, *arXiv preprint*, (2023), arXiv:2307.13640. https://doi.org/10.48550/arXiv.2307.13640

20. J. Hur, S. Roth, Optical flow estimation in the deep learning age, in *Modelling Human Motion: From Human Perception to Robot Design*, Springer, (2020), 119–140. https://doi.org/10.1007/978-3-030-46732-6_7

21. S. Shah, X. Xiang, Traditional and modern strategies for optical flow: An investigation, *SN Appl. Sci.*, **3** (2021), 1–14. https://doi.org/10.1007/s42452-021-04227-x

22. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in *IEEE Conference on Computer Vision and Pattern Recognition*, (2015), 3431–3440. https://doi.org/10.1109/cvpr.2015.7298965

23. A. A. Rafique, A. Jalal, K. Kim, Statistical multi-objects segmentation for indoor/outdoor scene detection and classification via depth images, in *2020 17th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, (2020), 271–276. https://doi.org/10.1109/IBCAST47879.2020.9044576

24. S. Civilibal, K. K. Cevik, A. Bozkurt, A deep learning approach for automatic detection, segmentation and classification of breast lesions from thermal images, *Expert Syst. Appl.*, **212** (2023), 118774. https://doi.org/10.1016/j.eswa.2022.118774

25. D. Yu, Q. Li, X. Wang, C. Xu, Y. Zhou, A cross-level spectral–spatial joint encode learning framework for imbalanced hyperspectral image classification, *IEEE Trans. Geosci. Remote Sens.*, **60** (2022), 1–17. https://doi.org/10.1109/TGRS.2022.3203980

26. A. F. Al-Battal, I. R. Lerman, T. Q. Nguyen, Object detection and tracking in ultrasound scans using an optical flow and semantic segmentation framework based on convolutional neural networks, in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2022), 1096–1100. https://doi.org/10.1109/ICASSP43922.2022.9747608

27. C. Xu, Q. Li, X. Jiang, D. Yu, Y. Zhou, Dual-space graph-based interaction network for RGB-thermal semantic segmentation in electric power scene, *IEEE Trans. Circuits Syst. Video Technol.*, **33** (2023), 1577–1592. https://doi.org/10.1109/TCSVT.2022.3216313

28. F. Sun, P. Ren, B. Yin, F. Wang, H. Li, CATNet: A cascaded and aggregated transformer network for RGB-D salient object detection, *IEEE Trans. Multimedia*, **2023** (2023), 1–14. https://doi.org/10.1109/TMM.2023.3294003

29. J. Li, K. Liu, Y. Hu, H. Zhang, A. A. Heidari, H. Chen, et al., Eres-UNet++: Liver CT image segmentation based on high-efficiency channel attention and Res-UNet++, *Comput. Biol. Med.*, **158** (2023), 106501. https://doi.org/10.1016/j.compbiomed.2022.106501

30. R. Raza, U. I. Bajwa, Y. Mehmood, M. W. Anwar, M. H. Jamal, dResU-Net: 3D deep residual U-Net based brain tumor segmentation from multimodal MRI, *Biomed. Signal Process. Control*, **79** (2023), 103861. https://doi.org/10.1016/j.bspc.2022.103861

31. S. Hou, T. Zhou, Y. Liu, P. Dang, H. Lu, H. Shi, Teeth U-Net: A segmentation model of dental panoramic X-ray images for context semantics and contrast enhancement, *Comput. Biol. Med.*, **152** (2023), 106296. https://doi.org/10.1016/j.compbiomed.2022.106296

32. H. Yang, D. Yang, CSwin-PNet: A CNN-Swin Transformer combined pyramid network for breast lesion segmentation in ultrasound images, *Expert Syst. Appl.*, **213** (2023), 119024. https://doi.org/10.1016/j.eswa.2022.119024

33. L. Willems, J. Vermeulen, A. Wiegerinck, S. Fekkes, M. Reijnen, M. Warle, et al., Construct validity and reproducibility of handheld ultrasound devices in carotid artery diameter measurement, *Ultrasound Med. Biol.*, **49** (2023), 866–874. https://doi.org/10.1016/j.ultrasmedbio.2022.11.013

34. B. C. Russell, A. Torralba, K. P. Murphy, W. T. Freeman, LabelMe: A database and web-based tool for image annotation, *Int. J. Comput. Vision*, **77** (2008), 157–173. https://doi.org/10.1007/s11263-007-0090-8

35. S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, et al., Adaptive histogram equalization and its variations, *Comput. Vision Graphics Image Proc.*, **39** (1987), 355–368. https://doi.org/10.1016/S0734-189X(87)80186-X

36. H. Xu, J. Zhang, J. Cai, H. Rezatofighi, D. Tao, GMFlow: Learning optical flow via global matching, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 8122–8130. https://doi.org/10.1109/CVPR52688.2022.00795

37. Z. Teed, J. Deng, RAFT: Recurrent all-pairs field transforms for optical flow, in *Computer Vision–ECCV 2020*, Springer, (2020), 402–419. https://doi.org/10.1007/978-3-030-58536-5_24

38. Z. Huang, X. Shi, C. Zhang, Q. Wang, K. C. Cheung, H. Qin, et al., FlowFormer: A transformer architecture for optical flow, in *Computer Vision–ECCV 2022*, Springer, (2022), 668–685. https://doi.org/10.1007/978-3-031-19790-1_40

39. K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, in *Computer Vision–ECCV 2016*, Springer, (2016), 630–645. https://doi.org/10.1007/978-3-319-46493-0_38

40. Z. Tian, C. Shen, H. Chen, T. He, FCOS: Fully convolutional one-stage object detection, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 9627–9636. https://doi.org/10.1109/ICCV.2019.00972

41. T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 936–944. https://doi.org/10.1109/CVPR.2017.106

42. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in *2016 IEEE Conference on Computer Vision and Pattern Recognitio (CVPR)*, (2016), 779–788. https://doi.org/10.1109/CVPR.2016.91

43. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in *2014 IIEEE Conference on Computer Vision and Pattern Recognitio (CVPR)*, (2014), 580–587. https://doi.org/10.1109/CVPR.2014.81

44. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in *Computer Vision–ECCV 2020*, Springer, (2020), 213–229. https://doi.org/10.1007/978-3-030-58452-8_13

45. B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, R. Girdhar, Masked-attention mask transformer for universal image segmentation, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 1290–1299. https://doi.org/10.1109/CVPR52688.2022.00135

46. T. Zhang, S. Wei, S. Ji, E2EC: An end-to-end contour-based method for high-quality high-speed instance segmentation, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 4433–4442. https://doi.org/10.1109/CVPR52688.2022.00440
47. K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, et al., MMDetection: Open MMLab detection toolbox and benchmark, *arXiv preprint*, (2019), arXiv:1906.07155. https://doi.org/10.48550/arXiv.1906.07155
48. K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 2980–2988. https://doi.org/10.1109/ICCV.2017.322
49. D. Bolya, C. Zhou, F. Xiao, Y. J. Lee, YOLACT: Real-Time instance segmentation, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 9156–9165. https://doi.org/10.1109/ICCV.2019.00925
50. Z. Tian, C. Shen, H. Chen, Conditional convolutions for instance segmentation, in *Computer Vision–ECCV 2020*, Springer, (2020), 282–298. https://doi.org/10.1007/978-3-030-58452-8_17
51. X. Wang, T. Kong, C. Shen, Y. Jiang, L. Li, SOLO: A simple framework for instance segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2022), 8587–8601. https://doi.org/10.1109/TPAMI.2021.3111116
52. A. Kirillov, Y. Wu, K. He, R. Girshick, PointRend: Image segmentation as rendering, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 9796–9805. https://doi.org/10.1109/CVPR42600.2020.00982
53. X. Wang, T. Kong, C. Shen, Y. Jiang, L. Li, SOLOv2: Dynamic and fast instance segmentation, *arXiv preprint*, (2020), arXiv:2003.10152. https://doi.org/10.48550/arXiv.2003.10152
54. Y. Fang, S. Yang, X. Wang, Y. Li, C. Fang, Y. Shan, et al., Instances as queries, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 6890–6899. https://doi.org/10.1109/ICCV48922.2021.00683
55. Z. Tian, C. Shen, X. Wang, H. Chen, BoxInst: High-performance instance segmentation with box annotations, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 5439–5448. https://doi.org/10.1109/CVPR46437.2021.00540
56. T. Cheng, X. Wang, S. Chen, W. Zhang, Q. Zhang, C. Huang, et al., Sparse instance activation for real-time instance segmentation, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 4423–4432. https://doi.org/10.1109/CVPR52688.2022.00439
57. C. Lyu, W. Zhang, H. Huang, Y. Zhou, Y. Wang, Y. Liu, et al., Rtmdet: An empirical study of designing real-time object detectors, *arXiv preprint*, (2022), arXiv:2212.07784. https://doi.org/10.48550/arXiv.2212.07784

AIMS Press

© 2024 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)